Data mining In UK higher education institutions: law and policy

Article (Accepted Version)

Data Mining in UK Higher Education Institutions: Law and Policy

Andres Guadamuz and Diane Cabell[*]

Abstract

This article explores some of the issues surrounding data mining in the UK's higher education institutions (HEIs). Data mining is understood as the computational analysis of data contained in a text or dataset in order to extract new knowledge from it. There are two main ways in which HEIs are involved with data mining: in the process of conducting research, and as producers of data. As consumers, HEIs may have restrictions on the manner in which they can conduct research given the fact that it is likely that content will be protected by intellectual property rights. As producers, HEIs are faced with increasing pressure to make publicly-funded research available to the public through institutional repositories and other similar open access schemes, but some of these do not set out reuse policies for data. The article concludes that if more research was made available with adequate licensing strategies, then the question of whether data mining research is legal would be moot.

Keywords: data mining, databases, copyright, database right, licensing, open, Creative Commons, open access.

1. Introduction

Data or text mining (hereafter called "content mining") is a process that uses software that looks for interesting or important patterns in data that might otherwise not be observed. An example might be combining a database of journal articles about ground water pollution with one of hospital admissions to detect a pollution-related pattern of disease breakout.

It is also a useful tool in commerce. A credit card company might detect a correlation between ticket purchases from a particular airline with purchases of certain types of automobiles and develop a

---

[*] Andres Guadamuz is a Senior Lecturer in Intellectual Property Law at the University of Sussex.

Diane Cabell is a Visiting Academic at the Oxford University's eResearch Center as well as Corporate Counsel for Creative Commons and Executive Director of iCommons Ltd.

This paper has been prepared for Wikipedia founder Jimmy Wales in advising the Universities and Science Minister David Willetts on the terms of access to the proposed Gateway to Research project. See http://bit.ly/Ry0FWU.

The authors would like to thank the anonymous reviewers for their invaluable suggestions. We would also like to thank the Kusuma Trust UK for its generous support of the iCommons Open Collaboration Research Project without which this paper would not be possible. The authors would also like to thank Dr. Abbe Brown, Senior Lecturer at the University of Aberdeen, Dr. Dinusha Mendis, Senior Lecturer at Bournemouth University, Dr. Prodromos Tsiavos, adviser on legal issues of open data in the Greek Prime Minister's e-Government Task Force and the Special Secretary for Digital Planning, and Diane Peters, Creative Commons General Counsel for their helpful input.

marketing program uniting appropriate vendors. One McKinsey report states that the utilization of 'big data' in the sphere of public data alone could create €250 billion annual value to Europe's economy.[1]

Content mining is increasingly performed through automated systems. Databases, particularly those produced by scientific research, are far too large to be scanned by the human eye. However, the right to mine data is not assured by the law in most jurisdictions, and even where it is, the terms of access to the majority of research publication databases deny permission to do so. One recent study indicated that obtaining permission to mine the thousands of articles appearing on a single subject from the publishers holding the rights to the works would require 62% of a researcher's time. Many content owners, including research institutions, have yet to develop any policy on content mining.[2]

Talking specifically about higher education institutions (HEIs), content mining is of great interest to them both as users – when investigators use it as a research tool– and as producers of knowledge. There are open questions in both situations faced by HEIs. From the user perspective, HEIs want to know if its staff can use content mining in their everyday research, particularly in data-heavy subjects. From the perspective of HEis as creators, they have to be able to provide it using adequate reuse policies.

The overreaching objective of this article is to try to answer the open questions in both academic aspects. From the user side it will identify the current law with regards to data mining in order to ascertain the main legal barriers for research purposes. Looking at HEIs as producers, the study will look at the increasing shift towards open access requirements, and therefore we will analyse institutional data reuse policies and licensing to see if they hinder in any way content mining. This will hopefully help HEIs in shaping their research policies both as users and creators of knowledge.

This objective may seem modest, but this is an important time in which to answer the questions posed by content mining. HEIs are increasingly involved in this type of research,[3] and the legal pitfalls and uncertainties may very well stifle innovations coming from this type of work. Similarly, UK HEIs are under growing pressure to work under a framework that favours open access publishing, particularly in providing access to basic scientific data that can be reused by other researchers. Adequate data reuse policies would help to ensure that researchers from other institutions could conduct content mining operations without fear of infringement. It is with this in mind that the second part of the paper there is such a strong emphasis on reuse policies and practices at HEIs. This is particularly important because, while there is a growing body of work dealing with "big data" from a legal perspective,[4] there has yet to be a study that narrows down the topic to UK HEIs.

2. Content mining

[1] McKinsey Global Institute, *Big Data: The next frontier for innovation, competition and productivity*, (2011).

[2] McDonald, *Value and benefits of text mining*, March 2012 at http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx.

[3] See for example a list of data mining projects at the University of East Anglia: http://bit.ly/1gXCXLY.

[4] See for example Borghi M and Karapapa S, *Copyright and Mass Digitization: A Cross-Jurisdictional Perspective*, Oxford: Oxford University Press (2013), and Mayer-Schönberger V and Cukier K, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston: Houghton Mifflin Harcourt (2013).

It is an undeniable fact that databases are growing in number and size.[5] This increase in data has prompted a change in the way in which we look at large datasets, as it becomes impossible for humans alone to sift through new knowledge. As a response to this challenge, computational technologies and techniques are increasingly used to retrieve and analyse data held in something called "knowledge discovery in databases" (KDD). Data mining is a subset of this branch of data analysis. While it may not be perfect, the mining analogy serves to explain roughly what content mining entails. Artificial intelligence agents sift through large amounts of data, eventually finding valuable information that was undiscovered before. Moreover, in large mining operations one sifts through large quantities of low-grade material in order to find something valuable.

As explained by Fayyad et al:

> *KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data.*

For the purposes of the present study, content mining is to be described as the extraction of data from large datasets to uncover previously unknown and potentially useful information.[6] While the field is relatively new, increased computing capabilities make the analysis of large datasets not only possible, but also useful. The applications for content mining range from the mundane to the transcendental. For example, studies have used text-mining techniques to explore social sentiment[7] and public opinion[8] through the analysis of social media. Other studies have been looking at the use of social media to survey health and disease occurrences, for example, by looking for the prevalence of mentions of influenza online.[9] More serious applications include the use of content mining in biology and medicine.[10]

The methods for extracting and analysing the data may be relevant for the legal questions that are the subject of this analysis. There are various types of content mining, for example, some look at anomalous records, or look for correlations and/or dependencies in the data. These techniques use different software and algorithms, so it is difficult to generalise for legal purposes. However, the statistical analysis usually associated with content mining requires access to the data, and the possibility of creating some form of remote copy for analysis purposes (although actual copies are

---

[5] Fayyad U, Piatetsky-Shapiro G, and Smyth P, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine* 37 (1996).

[6] Frawley WJ, Piatetsky-Shapiro G, and Matheus CJ, "Knowledge Discovery in Databases: An Overview", 13:3 *AI Magazine* 57 (1992).

[7] Pang B and Lee L, "Opinion Mining and Sentiment Analysis", 2:1 *Foundations and Trends in Information Retrieval* 1 (2008).

[8] O'Connor B et al, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (2010).

[9] Corley C et al, "Text and Structural Data Mining of Influenza Mentions in Web and Social Media", 7:2 *International Journal of Environmental Research and Public Health* 596 (2010).

[10] See for example Krallinger M, Valencia A and Hirschman L, "Linking genes to literature: text mining, information extraction, and retrieval applications for biology", 9:2 *Genome Biology* S8 (2008); and Ananiadou1 S, Kell DB, and Tsujii J, "Text Mining and its Potential Applications in Systems Biology" 24:12 *Trends in Biotechnology* 571 (2006).

not always necessary). Similarly, the analysis of the data tends to be aggregated and reused to produce tables, diagrams and histograms of the combined sets.[11]

It is difficult to generalise on what exactly is the method for content mining, as there are different algorithmic and model structures depending on the subject, the type of database, and the type of analysis being performed.[12] For the purpose of this study, it will be assumed that most content mining roughly follows these steps (Figure 1):

1. Individual content is created.
2. Content is placed into data set, repository or collection.
3. Miner gains access to the data.
4. Mining tools applied to the data set.
5. Analysis of the processed data.
6. New knowledge.[13]

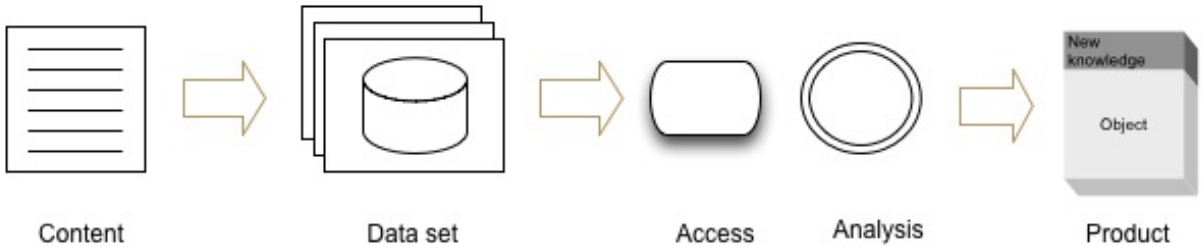

Figure 1. A typical content mining operation.

The key points from a legal perspective are stages 3 and 4. Researchers must be able to have access to the data in a format that is susceptible of analysis, for which it must be assumed that the content is either freely available, or the researcher has some form of licensing agreement allowing access. Then, there is the vital question of what operation is performed on the data. Is there copying of the entire content of the database? If not, what sort of operation is performed? Is there some form of retrieval of key data? Is the operation simply looking at patterns? What is the format of the new knowledge?

The answer to these questions may prove vital in answering the legality of content mining operations. In the interest of a general legal analysis, it will be assumed that there is actual copying of substantial sections of contents during the mining operation, although it is understood that this may not always be the case. It will also be assumed that the analysis operation means that the work has been extracted in the meaning of the database right, although this may also be open to interpretation.

---

[11] Han J and Kamber M, *Data Mining: Concepts and Techniques*, San Francisco, CA: Morgan Kaufmann Publishers (2000), p.16.

[12] Ibid, p.23.

[13] These steps are a simplified version of the processes described here: Korn N, Oppenheim C and Duncan C, *IPR and Licensing issues in Derived Data*, JISC report (2007), http://bit.ly/TEmtMX.

3. The law

Databases are protected in the UK through a variety of norms, and each may have a bearing on the legality of content mining.

3.1 Copyright

The data contained in databases can be protected under copyright law as a literary work. Section 3A of the Copyright, Designs and Patents Act 1988 (CDPA), defines a database as a collection of independent works which "are arranged in a systematic or methodical way", and "are individually accessible by electronic or other means". However, the threshold of originality in a database is quite high. Section 3A states that:

> *For the purposes of this Part a literary work consisting of a database is original if, and only if, by reason of the selection or arrangement of the contents of the database constitutes the author's own intellectual creation.*

This means that in UK copyright law the author's own "intellectual creation" is required in the selection and arrangement of the contents of a database, a mere gathering of data without meeting this requirement is not worthy of protection because it does not meet the originality test. There has now been extensive body of case law trying to define precisely what is meant by the phrase "own intellectual creation".[14] Of particular relevance to the issue of originality in databases is the case of *Bezpečnostní softwarová asociace*,[15] in which the Court of Justice of the European Union (CJEU) was asked to determine whether a graphic user interface (GUI)[16] in a computer program would be considered an author's own intellectual creation worthy of copyright protection. The Court decided that a graphic user interface is simply a manner in which a work can be user-friendly, and different source code and object code can have similar GUIs, so it is not part of a computer program.[17] However, the court found that the GUI could have copyright protection on its own right if it met the originality requirement; the problem being that many elements of a program are functional in nature, and therefore not worthy of protection. Similarly, it was determined that many such functional elements are simply not original enough because they are limited methods of implementing an idea, and therefore do not constitute an expression of the author's own intellectual creation.[18] The relevance of this case to the issue of databases is that in that situation we also encounter significant functional elements, such as the way in which the database is constructed and perform its function, and this is in some manner separate from the content itself of the database

Another UK case serves to illustrate the higher originality threshold in databases described above. In the English case of *Navitaire v Easyjet,*[19] Pumfrey J had to consider whether a computer-based database is a computer program or a database for copyright purposes, and interestingly found that the addition and removal of datasets, schemas and other structural changes to the arrangement of a

---

[14] Amongst these, see: Infopaq International C-5/08; Painer v Standard Verlags GmbH C-145/10; SAS Institute Inc v World Programming Ltd C-406/10, just to name a few.

[15] *Bezpečnostní softwarová asociace v Ministerstvo kultury* C-393/09.

[16] The way in which a program is displayed in a screen or interface.

[17] *Bezpečnostní,* paras 65-68.

[18] Ibid, paras 69-76.

[19] *Navitaire Inc v Easyjet Airline Co. & Anor* [2004] EWHC 1725 (Ch).

database were to be considered computer programs instead of databases in their own right. The meaning of this ruling for databases is that there would be a protection of the source code in the shape of a literal work, and not of the functional elements as such, which are an important and integral part of a database. The case spells out this dichotomy when Pomfrey J states clearly that "*Copyright protection for computer software is a given, but I do not feel that the courts should be astute to extend that protection into a region where only the functional effects of a program are in issue."*[20] While Navitaire is mostly cited in the context of software patents, it is highly relevant here because it can be understood as making the functional elements of a database largely irrelevant for the purpose of protection. A database is not only code, it is also the function that it serves, mostly in the shape of containing algorithms, search functions, and a functional syntax. If these are ignored, what we have left is the protection of the contents themselves, and of the software code that surrounds it.

Further the functional element found in databases, *Football DataCo,*[21] can be used to stress the fact that copyright in databases has a higher threshold. The case involved the fixture lists of football matches in the English and Scottish leagues, which are produced by a company called Football DataCo. Web aggregator Yahoo! copied these fixtures without paying licence fees, so Football DataCo sued them alleging that by doing so Yahoo! had infringed both copyright and its database rights. The Court of Appeal of England and Wales referred[22] the case to the CJEU, which decided that copyright can only be afforded to a database if its structure is the maker's own intellectual creation. This continued to set a bar high of not only originality, but of the originality required to have protection under copyright for a database. The CJEU opined that "the significant labour and skill required for setting up that database cannot as such justify such a protection if they do not express any originality in the selection or arrangement of the data which that database contains."[23]

Assuming copyright in the database exists, regardless of the high protection threshold, then the author would have the exclusive right to authorise use and reuse of the data, and any such unauthorised use would be a copyright infringement. Acts that infringe copyright might still fall under an exception or limitation, which in the UK take the shape of fair dealing. Only those acts listed under the CDPA can be considered exceptions. Section 50D does contain a fair dealing provision with regard to databases. It reads:

> *(1) It is not an infringement of copyright in a database for a person who has a right to use the database or any part of the database, (whether under a licence to do any of the acts restricted by the copyright in the database or otherwise) to do, in the exercise of that right, anything which is necessary for the purposes of access to and use of the contents of the database or of that part of the database.*

Unfortunately, this is a very narrow exception that is unlikely to cover the type of reuse of the information that is typical of content mining. Fair dealing in databases covers only those acts that are necessary to use the contents of the database, and in the strictest sense, one could argue that content mining is not a "necessary" use of the data, as the above exception seems to give permission on the basis of operational uses. Therefore, only functional uses could be considered non-infringing.

---

[20] At para 94.

[21] *Football Dataco Ltd and Others v Yahoo! UK Ltd and Others* C-604/10.

[22] [2010] EWCA Civ 1380.

[23] C-604/10 at para 46.

Similarly, content mining does not seem to fall under any other research-related fair dealing, as these also tend to be very narrow. For example, s29 CDPA states that:

> *(1) Fair dealing with a literary, dramatic, musical or artistic work for the purposes of research for a non-commercial purpose does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement.*
>
> *(1A) Fair dealing with a database for the purposes of research or private study does not infringe any copyright in the database provided that the source is indicated.[…]*
>
> *(1C) Fair dealing with a literary, dramatic, musical or artistic work for the purposes of private study does not infringe any copyright in the work.*

Any content mining operation that copies text would fall under this exception if it is for non-commercial purposes only, or if it is performed with the purpose of "private study". The definition clearly implies that content mining of medical texts by a pharmaceutical company looking for new drug treatment would clearly be an infringement, while content mining performed by an academic to do the same would find itself in more of a grey area. The problem with the research and private study exception is that, as Cornish points out, the courts have not been asked to ascertain how much can be taken, and what constitutes non-commercial use exactly.[24] The provisions can be interpreted in light of the InfoSoc Directive,[25] which in Art 5(b) contains a more comprehensive definition of what is to be considered as fair dealing for research; it reads:

> *…in respect of reproductions on any medium made by a natural person for private use and for ends that are neither directly nor indirectly commercial, on condition that the rightholders receive fair compensation which takes account of the application or non-application of technological measures referred to in Article 6 to the work or subject-matter concerned*.

It could be argued that academic research might fall under indirectly commercial use under some circumstances. Similarly, the request that rights holders should receive fair compensation denotes the restrictive interpretation given to the exception. Furthermore, content mining does not appear to fall under the exception for observing, studying and testing of computer programs (s 50BA).

The absence of a specific exception for content mining seems to indicate that if a database has copyright, most types of unauthorised content mining could be copyright infringement.

3.1.1 Temporary versus permanent copies

As it was stated earlier this paper is working under the assumption that in any content mining process there is actual copying involved. But what if this assumption is false, and the processing takes place by means of a temporary copy?

This is a vital question because transient or incidental copies that are part of an essential technological process are exempt of the reproduction right, in accordance to Art 5(1) of the InfoSoc Directive,[26] transposed into UK copyright law in s28A of the CDPA. This exemption can only occur if the transient copy is made to enable:

---

[24] Cornish WR Llewelyn D and Aplin T, *Intellectual Property: Patents, Copyright, Trade Marks & Allied Rights*, 8th ed, London: Sweet & Maxwell (2013), p.509.

[25] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

[26] Ibid.

(a)a transmission of the work in a network between third parties by an intermediary; or

(b)a lawful use of the work;

and which has no independent economic significance.[27]

While it is clear that content mining does not fulfil requirement (a), it could be argued that under some circumstances it might fall under (b), particularly if it has no commercial value. This might be the trickiest part for a content mining operation, if it is performed for pure research purposes, it might still have economic significance.

There are several cases[28] that might help to elucidate the application of the transient copy exception. The first is the CJEU case of *Infopaq International*,[29] where the Danish news-clipping service Infopaq International was taken to court by the Danish newspaper association Danske Dagblades Forening (DDF), over its reproduction of news cuttings for sale to its clients. The clipping process involved a data capture process consisting of scanning images of original articles, the translation of those images into text, and the creation of an 11-word snippet for sale to Infopaq's clients. While the CJEU admitted that some of the processes involved were transient, and therefore might be covered by the aforementioned exceptions, the fact that the copying process resulted in clippings that could be sold meant that the copying was too permanent to qualify for the transient copy exception.

After the first decision, the Danish court sent back the case for a second round of questions to clarify the reach of the meaning of transient copying in another cased known as *Infopaq II*.[30] The court expanded the definition by stating that acts of temporary reproduction must not have an independent economic significance in two ways, firstly, that the copying "does not enable the generation of an additional profit", and secondly, that "the acts of temporary reproduction do not lead to a modification of that work".[31]

The result in the first *Infopaq* case may not be entirely relevant to content mining, as the processing is not similar to that which occurs in data analysis, although it provides a useful delimitation of what constitutes copying. *Infopaq II* is considerably more useful to content mining because it sets a non-profit requirement for transient copying operations, but also that it requires that the reproduction must not lead to an adaptation. The first threshold might not be too important for HEIs, but it is almost certain that content mining would lead to modification of a work, as it is almost a requirement that the data analysis should result in some sort of new knowledge. The question is whether that amounts to a modification as implied in this case.

A more adequate test is the recent Supreme Court (SC) case of the *Copyright Licensing Agency (CLA) v Meltwater*[32] which has been exploring precisely the nature of temporary copying on the Internet by the means of indexing and caching, in a manner that is analogous with data mining operations.

---

[27] s28A CDPA.

[28] Besides the ones mentioned in this section, readers may want to look at *Football Association Premier League Ltd and Others v QC Leisure and Others* C-429/08.

[29] *Infopaq International A/S v Danske Dagblades Forening* C-5/08.

[30] *Infopaq International A/S v Danske Dagblades Forening* C-302/10.

[31] At para 54.

[32] *Public Relations Consultants Association Ltd v The Newspaper Licensing Agency Ltd & Ors* [2013] UKSC 18.

Meltwater News is a service that monitors newspaper websites and through the use of autonomous agents (also known as crawlers or spiders) it produces an index of all words present in those sites. This data is then sold to the members of the Public Relations Consultants Association (PRCA), and it allows PR professionals with advanced search tools on specific names and words in the news. CLA sued Meltwater and the PRCA for copyright infringement, as it had put in place a licensing scheme for media monitoring organisations. The defendants claimed that their actions were lawful because any copying was transient in nature. CLA won both the first instance and the appeal, and the case made it all the way to the Supreme Court.

The SC had to deal with the major issue of whether any copying performed by Meltwater had been temporary, and in doing so it produced an interesting discussion about the reaches of the transient copy exception. CLA had argued that the transient copy exception only applied to actions that would enable a transmission of the work. The SC did not buy this argument at all. Lord Sumption comments:

> *In the first place, it is clear from the Directive's recitals, and in particular from recital 33, that it was intended that the exception should "include acts which enable browsing as well as acts of caching to take place." Browsing is not part of the process of transmission. It is the use of an internet browser by an end-user to view web pages. It is by its very nature an end-user function.*[33]

The SC then argued that once it has been understood that the purpose of the transient copy exception is to enable users to view copyright material on the Internet, all other conditions must be read with that in mind. Lord Sumption is clear that there has never been an infringement of copyright for a person "merely to view or read an infringing article in physical form".[34] The SC gave a much wider interpretation to what constitutes lawful use than that given to in first instance and appeal stages. The SC argued that once it is accepted that temporary copies made for the purpose of browsing are made by an end-user, then it is acceptable to entertain the idea that Meltwater's actions may be exempt as temporary copies. The SC did not decide the case however, and referred the question of the transience of Internet communications to the CJEU. At the time of writing this has not been decided.

It will be interesting to keep an eye on these developments, but it might be better for content miners to assume that courts in the EU will consider their actions to be copying that does not fall under temporary copying exceptions. This is a big assumption to make, but the decisions from recent cases would bear that interpretation.

3.2 Database right

In addition to copyright protection for databases, the UK has implemented a sui generis right arising from the European Database Directive,[35] enacted in the UK through the Copyright and Rights in Databases Regulations 1997 (CRDR). It is important to point out that the database right exists regardless of the existence of copyright protection in the database, as the exclusive rights given to the database owner are separate to those arising from copyright.[36]

---

[33] At para 27.

[34] At para 36.

[35] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

[36] s13 CRDR.

The database right is an exclusive right given to the maker of a database,[37] which is defined as a collection of independent works, data or other materials that are arranged in a systematic or methodical way, and are individually accessible by electronic or other means.[38] The right exists if "there has been a substantial investment in obtaining, verifying or presenting the contents of the database".[39] The right subsists for 15 years from the completion of the same.[40] The right is infringed if a person without authorisation "extracts or re-utilises all or a substantial part of the contents of the database".[41] The right is also infringed after continuous extraction or re-utilisation of non-substantial parts of the database.[42] For the purpose of the CRDR, re-utilisation is understood as making the contents of the database available to the public by any means.[43]

The database right comes with a fair dealing provision stating that there is no infringement if a substantial part has been extracted[44] or re-utilised if:

> *(a) that part is extracted from the database by a person who is apart from this paragraph a lawful user of the database,*
>
> *(b) it is extracted for the purpose of illustration for teaching or research and not for any commercial purpose, and*
>
> *(c) the source is indicated.*[45]

It is clear that the database right, if it exists in a database, precludes many forms of unauthorised content mining operations. The fair dealing provision cited above applies only if the person performing the content mining is already a lawful user of the database, the operation is done with attribution, and for research-related non-commercial purposes. We encounter here the same problem about the lack of definition of what constitutes non-commercial use. It may be advisable to interpret this provision also in light of the InfoSoc Directive, as was done in the previous section with regards to copyright. This would mean that any direct or indirect commercial use might be infringing. For example, an academic who is funded by a pharmaceutical company for his research at the university might fall outside of what is permitted under fair dealing.

However, the CJEU delivered a set of decisions that watered down the database right by raising the bar of what databases can be said to meet the standard of protection. In 2004, the CJEU delivered a number of decisions clarifying the database right, of which one was a referral from an English court. In *British Horseracing Board v William Hill*[46], the CJEU was asked to determine whether the

---

[37] s14 CRDR.

[38] s6 CRDR.

[39] s13 CRDR.

[40] s17 CRDR.

[41] s16 CRDR.

[42] Ibid.

[43] Bently L and Sherman B, *Intellectual Property Law*, 3rd ed, Oxford: Oxford University Press (2008), p.303. However, this statement should be read in conjunction to the interpretation of re-utilisation set out in the Football Dataco cases described below.

[44] Or the continuous extraction of a non-substantial part as per s16 CRDR.

[45] s20 CRDR.

[46] *British Horseracing Board Ltd v William Hill Organization Ltd (BHB decision)* C-203/02.

collection of horse racing information obtained through a third party by the defendants was a database subject to the sui generis right. The betting agency William Hill obtained horse racing data by a licensing agreement with a third party, not with the British Horseracing Board, which created the data. While most of the case rested on the issue of whether there had been substantial extraction of data from the original, an important part of the decision was in regard to whether the database maker had incurred enough investment to warrant protection. Here the court decided that:

> *The expression 'investment in ... the ... verification ... of the contents' of a database in Article 7(1) of Directive 96/9 must be understood to refer to the resources used, with a view to ensuring the reliability of the information contained in that database, to monitor the accuracy of the materials collected when the database was created and during its operation. The resources used for verification during the stage of creation of materials which are subsequently collected in a database do not fall within that definition.*[47]

The above paragraph seems harsh, as in it the CJEU seems to seriously erode database protection by setting a high standard of protectable investment. The paragraph is particularly severe when it comes to the investment in verifying information that goes into a database. Here the CJEU further comments:

> *...although the search for data and the verification of their accuracy at the time a database is created do not require the maker of that database to use particular resources because the data are those he created and are available to him, the fact remains that the collection of those data, their systematic or methodical arrangement in the database, the organization of their individual accessibility and the verification of their accuracy through the operation of the database may require substantial investment in quantitative and/or qualitative terms within the meaning of Article7(1) of the Directive.*[48]

This means that the CJEU has not done away with verification altogether, it simply establishes high level of investment in all of those steps is required. As many commentators have noted, this significantly reduces the potential scope of the database right, as only those databases that meet the higher standard of investment are protected.[49]

A more recent series of cases involving football fixtures[50] help to elucidate the borders of the database right, and might be applicable to content mining. In the case of *Football Dataco v Yahoo!*,[51] several companies involved in the creation and management of football fixtures for the English and Scottish leagues sued the defendants, who were a number of persons and organisations using the fixtures without a licence, mostly for reporting and betting purposes. The claimants argued that the drawing of the fixtures required considerable skill and labour to create. The trial judge agreed with

---

[47] Ibid at para 31.

[48] Ibid at para 36.

[49] Davison MJ, Hugenholtz PB, "Football fixtures, horse races and spin-offs: the CJEU domesticates the database right", 3 *European Intellectual Property Review* (2005).

[50] For the sake of brevity we will not refer to the series of cases in *Football Dataco Ltd & Ors v Sportradar GmbH & Anor* [2011] EWCA.

[51] *Football Dataco and Others v Yahoo and Others* [2012] EUECJ C-604/10.

this view and declared that the fixtures were subject to copyright protection.[52] The Court of Appeal[53] tended to disagree with this interpretation, and examined the facts in light of Art 3 of the Database Directive, which establishes that a database will be subject to copyright protection if "by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation". Jacob LJ agreed that the making of the fixtures involved considerable judgement and skill, but wondered whether these were of the "right kind" for the purpose of Art 3. The case was therefore referred to the CJEU with two questions, of which the first is relevant to the current topic:

> *1. In Article 3(1) of Directive 96/9/EC on the legal protection of databases what is meant by "databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation" and in particular:*
>
> *(a) should the intellectual effort and skill of creating data be excluded?*
>
> *(b) does "selection or arrangement" include adding important significance to a pre-existing item of data (as in fixing the date of a football match);*
>
> *(c) does "author's own intellectual creation" require more than significant labour and skill from the author, if so what?*

The CJEU started by restating what should have been obvious from the start, that the copyright in Art 3(1), and the sui generis right, are both separate rights with different object and application. As such, the court held that the resources employed in the drafting of the fixture tables were not relevant for the purpose of assessing whether the database is eligible for copyright protection. The main element of consideration for copyright purposes is originality, and this is met when authors express their creative ability "in an original manner by making free and creative choices"[54] through the selection and arrangement of data. On the contrary, originality does not exist when the creation of the database "is dictated by technical considerations, rules or constraints which leave no room for creative freedom".[55] Given that the referring court had declared that the making of the football tables involved significant skill and judgement, the CJEU specifically stated that this alone does not warrant the existence of originality, and therefore did not meant that the database would be given copyright protection. It was therefore a question for the referring court to answer if football fixtures met the originality requirement that it set out. [56]

The result in *Football Dataco* is tangentially relevant for content mining in as much as it sets a very high threshold of protection for copyright in databases in accordance to the Database Directive. Content mining operations do not try to copy the database structure, they analyse the contents. As such, we should revert to what the law says, and here Art 10(2) of the Trade-Related Aspects of Intellectual Property Rights (TRIPS) agreement, which states that the a database may be protected by copyright as a compilation by reason of the "selection or arrangement of their contents". Therefore, the discussion in the previous section would apply.

---

[52] *Football Dataco Ltd & Ors v Brittens Pools Ltd & Ors* [2010] EWHC 841.

[53] *Football Dataco Ltd & Ors v Yahoo! UK Ltd & Ors* [2010] EWCA Civ 1380.

[54] C-604/10, at para 38.

[55] Ibid, at para 39.

[56] Further analysis of the case can be found here: P Virtanen, "Football DataCo v Yahoo! The ECJ interprets the Database Directive", 9:2 *SCRIPTed* 258 (2012); and D Rose and N O'Sullivan, "Football Dataco v Yahoo! Implications of the ECJ judgment", 7:11 *Journal of Intellectual Property Law & Practice* 792 (2012).

From the results in the CJEU cases cited, it is increasingly likely that the database right has not met the initial expectations for which it was created. The European Commission conducted a review of the impact of the new right, and found that it had no effect whatsoever in fostering the creation of a new sector in the European economy. In 1996, the United States (which provides no sui generis database protection) had the largest share of the global database market, with 56%, while European share was 22%. While this share increased between 1996 and 2001, it had dropped again to 24% by 2004, while the U.S. share went back to its previous levels.[57] This is strong indication that the sui generis right did not have any noticeable effect in strengthening the European database market. In an indicting comment on policy based on lobbying and guesswork, the Commission's report said:

> *Nevertheless, as the figures discussed below demonstrate, there has been a considerable growth in database production in the US, whereas, in the EU, the introduction of "sui generis" protection appears to have had the opposite effect. With respect to "non-original" databases, the assumption that more and more layers of IP protection means more innovation and growth appears not to hold up.*[58]

Despite this, there are no plans to scrap the sui-generis right. Institutions involved in content mining in the UK and Europe will still have to take into consideration the Database Directive, although it seems evident that copyright is more of a concern than database right, as both the threshold of protection for the database itself, as well as what can be considered as protected data extraction have become quite high.

3.3 Public Sector Information

Academic institutions are major producers of research data. As most HEIs in the UK receive public funds in one way or another, it is necessary to cover the relevant norms that rule the use and reuse of public sector data. The regime in place was enacted by the 2003 Public Sector Information (PSI) Directive,[59] which has been implemented in the UK in the Re-Use of Public Sector Information Regulations 2005.[60] The purpose of the PSI system is to encourage the reuse of public sector information. Although neither the Directive nor the Regulations require public sector organisations to make documents available to the public, if they do so it should be in line with the notions of transparency, fairness and consistency.

The PSI Regulations establish an exhaustive list of institutions that are considered public sector bodies and therefore covered by the legislation. Educational institutions are specifically exempted from the Regulations in s 5(3)(b), which reads:

> *These Regulations do not apply to documents held by— […]*
>
> *(b) educational and research establishments, such as schools, universities, archives, libraries, and research facilities including organisations established for the transfer of research results;*

---

[57] European Commission, *First Evaluation of Directive 96/9/EC on the Legal Protection of Databases*, DG Internal Market Working Paper, http://is.gd/DsY3XV.

[58] Ibid, p.24.

[59] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. The Directive has recently been amended by the Directive 2013/37/EU.

[60] The Re-use of Public Sector Information Regulations 2005, SI No. 1515.

This exclusion is somewhat unfortunate because an important part of the UK's strategy has been the creation of a unified licensing scheme for public sector information, more of which will be covered below.

4. Open Access Policies

The UK is fast becoming one of the most forward-looking countries with regard to opening access to research, in part thanks to a shift in policy from funding bodies in favour of wider access to research, but also due to growing government pressure in that respect. The rise of open access[61] in higher education institutions is of great importance for content mining as it can free up databases and other resources to analytical exercises. This is particularly relevant because, as we have seen before, these works may be restricted either by copyright or by the database right.

Significant pressure to make research more openly available has come from investigators themselves, with prominent academic voices coming out in favour of open access.[62] One such example is the Manchester Manifesto,[63] a document drafted by UK and European scientists trying to answer the question "who owns science?" They conclude that:

> *Scientific information, freely and openly communicated, adds to the body of knowledge and understanding upon which the progress of humanity depends. Information must remain available to science and this depends on open communication and dissemination of information, including that used in innovation.*

Another valuable pillar in the success of open access has been the fact that funding bodies are increasingly requiring that any research they support financially must be released at some point to the public, be it via institutional repositories, self-publishing, or through other similar means. The Wellcome Trust has enacted an Open Access Policy which makes it clear that, while it expects funded research to be published in peer-reviewed journals, it also requires that such works should eventually be made available to the public for free through PubMedCentral UK[64] within six months of publication.[65] Similarly, Research Councils UK, the partnership of the seven higher education funding research councils, has also established an updated open access policy[66] which states that all publicly-funded research must be published in an open access journal that allows "immediate and unrestricted access to the publisher's final version of the paper". If the journal does not offer such an option, then the work must be published in a journal that allows the work to be placed in other repositories "without restrictions on non-commercial re-use and within a defined period". Such

---

[61] It is assumed that the reader is already familiar with open access. If that is not the case, the Berlin Declaration on Open Access defines it as "a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community. […] Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material." See: http://is.gd/HTZLr6.

[62] Mathematician Tim Gower boycott against Elsevier; Mark Walport or other signatories to Bethesda Statement on Open Access Publishing

[63] Addison T et al, *The Manchester Manifesto,* Institute for Science, Ethics and Innovation (2009).

[64] Soon to be Europe PubMed Central.

[65] See the policy here: http://is.gd/rHhQM9.

[66] RCUK's 2012 policy version can be found here: http://is.gd/xbjUDv.

clear and unequivocal statements in support of open access are transforming scientific publishing, and allow more works to be accessible for mining.

The UK government itself has also been directly responsible for encouraging wider adoption of open access. One of the main drivers of this push has been the Joint Information Systems Committee (JISC), which is an independent quasi-autonomous non-governmental organisation (QUANGO) supported by the main national higher education funding councils and by the Department for Employment and Learning. Its main role has been to support and finance internal and external projects related to all aspects of information management in education, including projects on digital repositories, archives, content mining, preservation, metadata, standards, and interoperability. In exercising this function, JISC has produced a considerable number of reports in favour of open access,[67] but it has also created a substantial infrastructure that provides tools necessary for open access.

An important part of the work of JISC when it comes to open access has been to promote and encourage researchers in HEIs to upload content to institutional repositories. Needless to say, this is a vital part of any open access strategy. Besides having published guides on how to promote the adoption and use of repositories,[68] JISC has funded projects that try to find ways in which to encourage open access.[69]

In November 2010, the government commissioned an independent review on how intellectual property supports growth and innovation. The Hargreaves Review of Intellectual Property[70] produced a series of interesting and balanced recommendations. The study specifically mentions text mining as a subject that requires a new exception in copyright. The Review states:

> *Text mining is one current example of a new technology which copyright should not inhibit, but does. It appears that the current non-commercial research "Fair Dealing" exception in UK law will not cover use of these tools under the current interpretation of "Fair Dealing". In any event text mining of databases is often excluded by the contract for accessing the database. The Government should introduce a UK exception in the interim under the non-commercial research heading to allow use of analytics for non-commercial use, as in the malaria example above, as well as promoting at EU level an exception to support text mining and data analytics for commercial use.*[71]

The current UK government administration has indicated its support for Hargreaves' recommendations in the belief that it will not only stimulate scientific research but will also enable greater commercialization of UK know-how. This potential was also recognized in the latest and most comprehensive review on open access, the Report of the Working Group on Expanding Access to Published Research Findings (Finch Report).[72] The group was established by the Minister for

---

[67] For some reports, see: http://is.gd/NKvBIb.

[68] See for example: http://bit.ly/NPPzI2.

[69] See for example: Proudfoot RE et al, *JISC Final Report: IncReASe (Increasing Repository Content through Automation and Services)*, White Rose Consortium (2009).

[70] Intellectual Property Office, *Digital Opportunity: A Review of Intellectual Property and Growth*, (2011), http://www.ipo.gov.uk/ipreview.htm.

[71] Ibid, para 5.26.

[72] *Accessibility, sustainability, excellence: how to expand access to research publications*. Report of the Working Group on Expanding Access to Published Research Findings: http://is.gd/91tsKb.

Universities and Science in the context of the Research Innovation Network, and was tasked with advising the government on it policies with regards to scientific research. Although the Report does not study content mining in depth or suggest any other solutions beyond those of the Hargreaves Review, the report comments:

> *Related to such moves has been a growth of interest in exploiting the potential of text-mining tools to analyse and process the information contained in collections or corpora of journal articles and other documents in order to extract relevant information, to manipulate it, and to generate new information. The use of such techniques is not yet widespread, not least because arrangements for making publications available for text mining can be complex, and because the entry costs are high for those who lack the necessary technical skills. But text mining offers considerable potential to increase the efficiency, effectiveness and quality of research, to unlock hidden information, and to develop new knowledge.*[73]

The Finch Report came out strongly in favour of open access as a matter of government policy, encouraging OA publishing through article processing or publishing charges (APC)[74] whereby the expense of publication in an open access journal is borne by the grantee research institution, whenever there have been public funds have been used in the research. Similarly, it advises that an effective public policy towards open access should be accompanied by an effort to "minimise restrictions on the rights of use and reuse, especially for non-commercial purposes, and on the ability to use the latest tools and services to organise and manipulate text and other content".[75] Although Finch's preference for the author-pays model (so-called "gold" open access as opposed to the "green" OA method which allows authors to self-publish the work in any open access repository) has prompted some criticism,[76] there can be little doubt that the above constitutes a fundamental shift in favour of future access to research, including access to reuse by content mining.

Even more encouraging is the announcement by the government that it will be implementing the Finch Report's recommendations. Furthermore, they have guaranteed that all future research funded by public money will be available without restrictions anywhere in the world.[77] Finally, open access advocates have started to campaign in earnest in favour of content mining of scholarly publications. In a recent article, molecular scientist and OA expert Peter Murray-Rust formulated the concept of "open content mining", defining it as:

> *… the unrestricted right of subscribers to extract, process and republish content manually or by machine in whatever form (text, diagrams, images, data, audio, video, etc.) without prior specific permissions and subject only to community norms of responsible behaviour in the electronic age.*[78]

---

[73] Ibid, para 3.19.

[74] Various terms are used to define this work

[75] Ibid, p.7.

[76] Ayris P, "Why panning for gold may be detrimental to open access research", *The Guardian* (23 July 2012), http://is.gd/uscUS3.

[77] Sample I, "Free access to British scientific research within two years", *The Guardian* (15 July 2012), http://is.gd/yOCTus.

[78] Murray-Rust P, "The Right to Read Is the Right to Mine", *Open Knowledge Foundation Blog* (June 1, 2012), http://bit.ly/O75Rwd.

In the article he proposes three main principles governing open content mining. These are:

1. *Right of Legitimate Accessors to Mine.* There should be no objection to automated analysis of published works in the interest of research.
2. *Lightweight Processing Terms and Conditions.* Licensing and other terms and conditions should not restrict mining.
3. *Use.* Researchers should be able to publish and disseminate the result of their analysis.

These principles are a sign of the growing importance of content mining, but are also a welcome addition to the intellectual and ethical push towards more open research environment. Until the open access government recommendations are fully implemented and assuming that a database is protected by copyright and/or the database right, then content mining can be performed legally only with adequate permission to do so.

5. Content mining and licensing

Given the emphasis given by both UK government and funding bodies towards open access, HEIs should have in place policies that will facilitate sharing. This is to be done usually through institutional policies that will empower third parties to have access to research that has been performed by its staff. Assuming that such open access policies will cover content mining, the next sections will explore existing policies to test if these are actually geared towards allowing analytical data retrieval.

This is where the terms and conditions governing data use and reuse require careful analysis. If we are thinking of higher education data, it often is held in a repository or archive of some sort. In order to facilitate access to data, these repositories should be released under some open access licence, or at least with an explicit share and reuse policy.

It will be assumed that readers are familiar with open licensing schemes.[79] In the context of data and research held in UK-based repositories, the following are the most important:

- Creative Commons: These are licences with the aim of promoting science and the arts by making it easier for authors and creators to offer a flexible range of protections and freedoms to users of their works. It counters the "all rights reserved" tradition associated with copyright by introduction a set of licences in which authors keep only "some rights reserved". These licences range from dedicating the work straight to the public domain, to more narrow licences with several restrictions.
- Open Data Commons (ODC): The Open Data Commons is a set of licences and dedications created by the Open Knowledge Foundation (OKF) that are specifically directed towards protecting databases. The ODC suite includes the Open Database Licence (ODbL),[80] the

[79] Some works dealing with open licences include: Liang L, *Guide to Open Content Licenses*, Rotterdam: Piet Zwart Institute (2004); Guadamuz Gonzalez A, "Open Science: Open Source Licences for Scientific Research", 7:2 *North Carolina Journal of Law and Technology* 321 (2006); and Dusollier S, "The Master's Tools V the Master's House: Creative Commons V Copyright", 29:3 *Columbia Journal of Law & the Arts* 271 (2007).

[80] Full text here: http://opendatacommons.org/licenses/odbl/1.0/.

Open Data Commons Attribution License,[81] and the Open Data Commons Public Domain Dedication and License (PDDL).[82]

- UK Government Licensing Framework: As part of the framework arising from the PSI Directive and PSI Regulations, the UK government has been heavily involved in releasing datasets to the public by offering data through its own data portal called Data.gov.uk.[83] Parts of these efforts have been to create specific licences for public sector data: the Open Government Licence[84] and the Non-Commercial Government Licence.[85] The licences cover both copyright and database right works, and allow the user to copy, publish, distribute, adapt and combine the information.

These three licensing suites would be the logical solution for HEIs interested in making research accessible to the public. Before looking at higher education institutional practice in detail, it is useful to know which open licences (if any) are prevalent in the wider open data scene. It is difficult at present to take a complete snapshot of licence usage and adoption, but there are some important pointers that may give an indication of the types of licences used to protect data.

The data.gov.uk repository is a good starting point because it offers daily metadata for each hosted dataset. As of 31 July 2012, the site listed 11,720 individual metadata records, of which 9,898 (84.4%) are licensed with the Open Government Licence; the rest are mostly not specified or have no licence metadata attached, and only a minority (less that 1%) use other licences. This is an impressive result, but not really surprising when one takes into account that the terms and conditions of the site clearly specify that:

> *The data and information available through www.data.gov.uk are available under terms described in the "licence" or "constraints" field of individual dataset records (meta-data). Except where otherwise noted this is the Open Government Licence.*
>
> *All dataset records (meta-data) published on www.data.gov.uk are licensed under the Open Government Licence.*

The above is a good indication that a clear set of licensing instructions can seriously increase specific licence adoption within an archive. Contrast that with the information gathered by a recent survey of databases in the OKF's own Data Hub catalogue.[86] This site offers no instructions to would-be licensors other than the fact that the site's metadata is licensed with the Open Database Licence. Of the 4,004 entries in that repository, an astounding 50% do not have any specific licence attached to them. This is surprising as the site favours open data, so one would expect a much higher level of open data sophistication. Of the datasets released with a licence, 31% used some form of CC licence, while only 11% used an Open Data Commons license. Only a minority used some form of UK government licence like the Open Government Commons (Figure 2).[87]

---

[81] Full text here: http://opendatacommons.org/licenses/by/.

[82] Full text here: http://opendatacommons.org/licenses/pddl/.

[83] http://data.gov.uk.

[84] Full text here: http://www.nationalarchives.gov.uk/doc/open-government-licence/.

[85] Full text here: http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/.

[86] Miller P, "Thinking about Open Data, with a little help from the Data Hub", *Cloud of Data* (31 July, 2012), http://bit.ly/MZG5vN.

[87] Ibid.

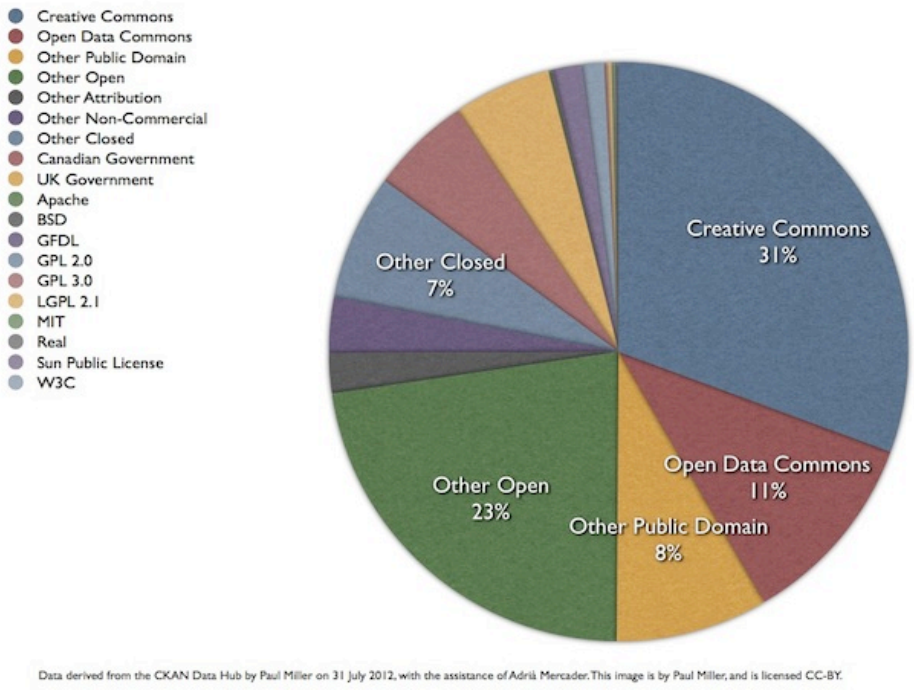

Figure 2. Types of licence used in Data Hub datasets.

This is an interesting finding for many reasons. Firstly, the current versions of Creative Commons licences are not specifically designed to work with the database right, so the datasets licensed under it may only cover the copyright element. Secondly, some of the other licences in use are not only not directed towards protecting databases, they are specifically software licences: e.g. the Apache Public License, the General Public License (GPL) in its various forms, and the Berkeley Software Distribution (BSD), just to name a few. This indicates that developers, owners and database makers in general are either not aware of other licensing choices, or they are aware of the existing licences and choose specific solutions because they are tailored to their needs. What seems clear is that licence choice is fragmented outside of the core UK government datasets, and this is not a favourable practice for potential content mining operations, as will be seen below.

Any attempt to measure open licence adoption may seem like an academic exercise, an attempt to distinguish between different flavours of the same thing. However, licence choice has very important consequences to reuse of content, as one licence may impose conditions that make it incompatible with other licence clauses used downstream. This is relevant particularly when dealing with collections, databases and other types of collective works. Incompatible licences could make it difficult to reuse and aggregate content from various sources.[88] This is one reason why CC licences remain very popular due to their high visibility and name recognition, as a strategist

[88] For examples of problems with licence incompatibility in open source software, see: Rosen LE, *Open Source Licensing: Software Freedom and Intellectual Property Law*, Upper Saddle River, N.J.: Prentice Hall PTR (2004), p. 267.

interviewed in a JISC report commented, "it's got to be CC [Creative Commons] or we're not using it. Because that just removes all the complexities."[89]

To illustrate this point, imagine a content mining project that gathers content from two different archives, one that uses a Creative Commons BY-NC-SA licence, and another one that uses the ODbL. At the time of writing, these licences are incompatible with each other because the ShareAlike element in CC licences only permits the user to distribute modified works under "the terms of this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License".[90] The ODbL contains a broader ShareAlike definition that allows the redistribution of adaptations with a "compatible license", but there is no list of compatible licences included, so in theory, both licences require derivatives to be published with their own terms. Furthermore, the NonCommercial element in the CC licence would also make it incompatible with the ODbL.

For the time being, potential users of incompatible content have the option of trying to gain permission to use another licence from the licensor. While this is cumbersome, it decreases legal issues arising from licence choice. It is true that many licensing institutions may not be aware of possible licence incompatibility, and may not even attempt to pursue a licence breach for the use of an incompatible licence. Nonetheless, wilful infringement is never recommended.

6. Higher education repositories

6.1 What is a repository?

The aforementioned strong institutional push towards open access from the UK government and important funding bodies has had a clear impact in higher education institutions. One of the most visible effects is the growth in institutional digital archive facilities, otherwise known as repositories, where academics and researchers can upload their own work in order to make it available to the public or the institution can have dedicated staff uploading, updating and maintaining such data. Technically, a repository is not the same as a mere online collection of works.[91] JISC defines digital repositories in the following manner:

> *A digital repository is a managed, persistent way of making research, learning and teaching content with continuing value discoverable and accessible. Repositories can be subject or institutional in their focus. Putting content into an institutional repository enables staff and institutions to manage and preserve it, and therefore derive maximum value from it. A repository can support research, learning, and administrative processes. They are commonly used for open access research outputs.*[92]

---

[89] White D and Manton M, *Open Educational Resources: The Value of Reuse in Higher Education*, JISC Report (2011), http://bit.ly/PwT3iR.

[90] s 4 b) CC BY-NC-SA 3.0.

[91] Heery R, *Digital Repositories Review*, Report for the United Kingdom Office for Library and Information Networking (2005).

[92] JISC, *Digital Repositories,* (2012), http://www.jisc.ac.uk/whatwedo/topics/digitalrepositories.aspx.

It is possible to classify repositories based on the type of submission. Some institutions have all-purpose repositories[93] where institutional content is stored; others have separate sites for theses, published articles and working papers,[94] while some institutions have subject-specific repositories.[95]

With regards to content mining, it is important both to be able to access the contents of a repository and to have the appropriate permission to reuse the content afterwards. Being technically accessible is precisely why it is important to talk about repositories and not just about some institutional website, as repositories tend to be published with software that will facilitate data searches and information retrieval. This criterion can only be met with an adequate technical infrastructure in place, preferably one that makes it easy not only to upload but also to search and access content. This is best accomplished if the information is stored with standard formats and in compliance with metadata standards.[96]

Because of the favourable policies outlined earlier, considerable investment has been made to support repository infrastructure both at the technical and logistic level. This has resulted in a technically favourable environment for content mining within the UK's higher education repositories. JISC in particular has been at the forefront of funding and supporting the development of institutional repositories. The result of such funding is a wealth of technical tools that allow ease-of-access to repository data.

6.2 Repository policies

It is evident that technical standards and tools are highly developed, but unfortunately the same cannot be said for the intellectual property issues surrounding repositories. While the open access ethos is on the rise, and the quality of content and database standard licences is also increasing, repositories do not always have clear policies on use and reuse of data and metadata. We conducted a survey of various aggregated data and of individual repositories, which produced relatively poor policy implementation.

There are several types of policies that can govern a repository. A report from the Data Information Specialists Committee-UK (DISC-UK)[97] describes the following types of policies:

- **Metadata policy:** for the information that describes items in the repository.
- **Data access and reuse policy:** for the items contained in the repository; this includes full-text works and other full data items.
- **Submission policy:** concerning various issues such as the identity of depositors, access, quality of content, formats, and most importantly for the purpose of this study, copyright policy.

---

[93] For an example see TeesRep, the Teeside University repository: http://tees.openrepository.com/tees/.

[94] The University of Birmingham has separate sites for theses (etheses.bham.ac.uk), published articles (eprints.bham.ac.uk), and working papers (epapers.bham.ac.uk).

[95] See the Electronic Gateway for Icelandic Literature at the University of Nottingham (www.egil.nottingham.ac.uk), and the First World War Poetry Digital Archive (www.oucs.ox.ac.uk/ww1lit) at Oxford.

[96] Ibid, p.18.

[97] Green A, MacDonald S and Rice R, *Policy-making for Research Data in Repositories: A Guide*, Report from the Data Information Specialists Committee-UK (2009), http://www.disc-uk.org/docs/guide.pdf.

- **Preservation policy:** concerns long-term issues, such as data sharing and archiving.

These four core types of policies reflect the highly complicated set of legal issues governing repositories. The IP aspects on their own are complex, as one must take into account the competing interests and needs of funders, researchers, students, and university departments. It is rare to find an institution-wide IP policy that covers all of the above parties and types of work.[98]

Researching the user terms and conditions of institutional repositories is a difficult endeavour because of the lack of clarity, and in many instances, the complete absence of policies and terms of use. We visited all of the sites linked to in the JISC-funded SHERPA institutional repository list,[99] looking for any indication of clear terms of access and reuse. Most sites visited had a submission copyright policy in place, so the terms and conditions were centred on providing an introduction to copyright for authors. In most sites, the policies were geared towards education and avoiding the submission of papers where the author did not have copyright in the work, and therefore were designed to minimise the institution's liability.[100] This is evidenced by the presence on several sites of procedures for removal ("take down") of copyright infringing content.[101] In some instances, the absence of key policies appears to be due to the use of technology that makes it difficult to present other documents besides the actual archive. For example, several institutions use DSpace software, which has a limited user interface that may discourage the inclusion of additional documentation.[102]

Of the 192 HEIs listed in the SHERPA institutional repository list, only 53 institutions had a publicly accessible repository, so we used those as a representative sample for analysis purposes. Of those 53 institutional repositories visited, 45 (84%) had some sort of copyright policy, but as stated before, these were mostly for submission purposes. In fact, of the total visited, only 20 sites (37%) had clear, easy-to-access and unambiguous data reuse policies. The sample indicates that while copyright awareness is high, there is still a long way to go towards converting that awareness into reuse policies.

It must be said that, where present, many institutions offer good submission practices, attempting to ensure that the database contents themselves are not infringing copyright. The University of Leicester has a good example of a concise set of guidelines to that effect:[103]

1. *Items may only be deposited by accredited members of the institution, or their delegated agents*
2. *Authors may only submit their own work for archiving*
3. *Eligible depositors must deposit full texts of all their publications, although they may delay making them publicly visible to comply with publishers' embargos*

---

[98] A good example of an institution that takes a holistic approach to IP is Oxford, see; http://www.admin.ox.ac.uk/rso/ip/.

[99] http://www.sherpa.ac.uk/guidance/instcontacts.html.

[100] See for example the Bristol Repository of Scholarly Eprints (ROSE), http://bit.ly/1jiuyt5, or the Anglia Ruskin Research Online user guide: http://libweb.anglia.ac.uk/academic/files/ARROguide.pdf.

[101] For example, see the Robert Gordon University policies: http://is.gd/og4LCA.

[102] E.g University of Hartfordshire and University of Edinburgh. The exception to this rule is the University of Leicester, which uses Dspace and has user guidelines: https://lra.le.ac.uk/.

[103] http://www2.le.ac.uk/library/about/policies/lra-policies.

4. *The administrator only vets items for the eligibility of authors/depositors, and relevance to the scope of Leicester Research Archive*
5. *The validity and authenticity of the content of submissions is the sole responsibility of the depositor*
6. *Items can be deposited at any time, but will not be made publicly visible until any publishers' or funders' embargo period has expired*
7. *Any copyright violations are entirely the responsibility of the authors/depositors*
8. *If Leicester Research Archive receives proof of copyright violation, the relevant item will be removed immediately.*

Regarding submission polices, we did not find in any of the repositories any example of further granularity in the terms and conditions with regards to the origin of the work. As stated above, repositories tend to be classed as general, thesis, published article and working paper. As such, there is no indication of the source of funding, i.e. whether the funding comes from private enterprises, or from government sources. This lack of distinction simplifies that policy; however the same repository may include works that are subject to conflicting rights regimes.

The Directory of Open Access Repositories (OpenDOAR)[104] contains a considerably more comprehensive list of 207 repositories in the UK. The divergence with the SHERPA list can be explained by the fact that the OpenDOAR is more up to date, but that it also lists archives belonging to non-HEIs as well as various institutions that have multiple repositories (e.g. the University of Southampton hosts 11 separate ones). As stated above, many other institutions maintain separate archives for theses and for academic papers.

The OpenDOAR has conducted a thorough survey of all of the repositories listed, and its figures are similar to our sample. They look at reuse policy for both metadata and data, as many websites have different policies for each.

For metadata, 61% of UK repositories have either unknown or undefined metadata policies. Of those with one in place, 10.6% allow for commercial use, while 28.4% allow reuse only for non-profit purposes (Figure 3).

[104] http://www.opendoar.org.

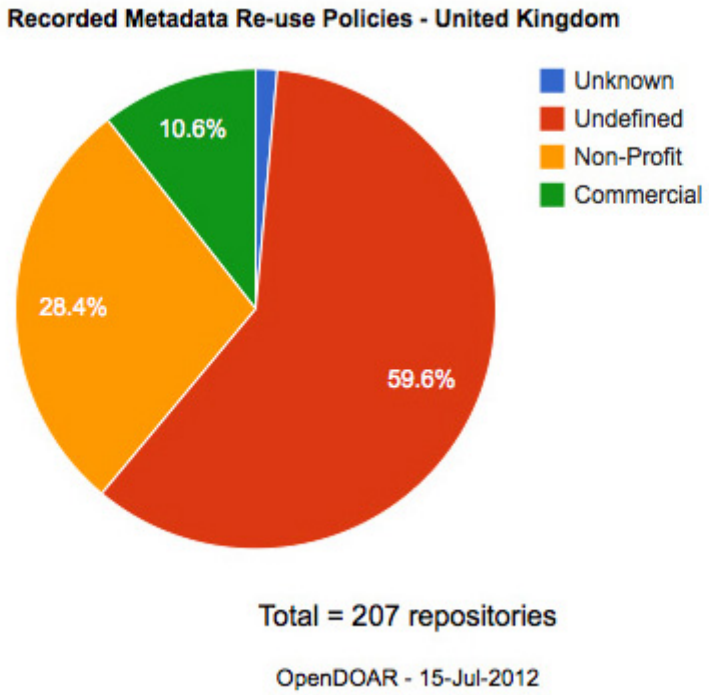

Figure 3. Recorded metadata re-use policies UK.[105]

For full-text data reuse, the OpenDOAR survey found that 57.7% of sites had an unknown, undefined or unclear policy in place. 18.8% had policies in which the rights varied for the reuse of full data items, and 13.9% only allowed reuse for non-profit purposes. Interestingly, the survey found that 8.2% of sites did not allow full-text indexing of sites by mechanical means through the existence of a No Robots file (norobots.txt). This operates as a *de facto* prohibition for reuse of data, even if such a restriction is not intended (Figure 4).

[105] http://bit.ly/N3xHMW.

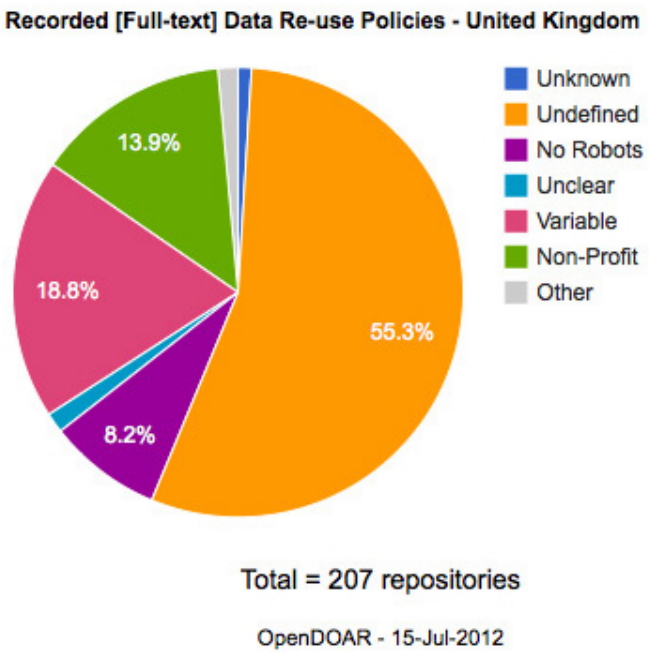

Figure 4. Recorded full-text data re-use policies UK.[106]

Both sets of statistics make for some worrying reading, as it is clear that even when available, the range of rights and restrictions on offer is too varied. When it comes to licensing, it could be said that less is more, and it would be desirable that one set of terms and conditions should prevail in one way or another, much as it does in the sample of databases licensed under the data.gov.uk site.

The source of the problem may come from the fact that these institutional repositories are not choosing their policies in a strategic manner due to the lack of harmonisation of licensing tools. Some sites are clearly using ad hoc policies,[107] while a few sites visited choose to use Creative Commons for reuse.[108] As stated before, these choices may not be compatible with databases; similarly, the reused materials from sites using CC licences incompatible with each other mean that those contents cannot be mixed without obtaining permission.

Most sites with reuse guidelines in place seem to be using the OpenDOAR Policy Tool. This is an application which generates text for five different types of policy: Metadata, Data, Content, Submission and Preservation. In each one of these fields, the institution chooses between a set of options to produce a page that can then be included in the repository. These options can be quite complex, for metadata alone users select between 10 variables, and for data there are 30 fields where selection is available. This goes a long way to explaining the statistics shown above, as it is clear that repositories are spoiled for choice. The disadvantage of this situation is that it creates

[106] http://bit.ly/N3xGID.

[107] See for example Aberystwyth University: http://bit.ly/1jiuhX5; and the University of St. Andrews: http://bit.ly/TJMGMU.

[108] Imperial College uses the generic BY-NC-ND 3.0 http://bit.ly/N3zQrY; while the Open University uses CC BY-NC-SA 2.0 England & Wales, see: http://bit.ly/N3zDF8.

interoperability issues if one wishes to reuse data from various different datasets, as some of the elements of choice are incompatible with one another.[109]

Nonetheless, the OpenDOAR Policy Tool produces some clear policy text for both metadata and data. Take the example of the metadata policy for the Nottingham ePrints repository,[110] which is typical of many other sites:

> *Metadata Policy for information describing items in the repository*
>
> 1. *Anyone may access the metadata free of charge.*
> 2. *The metadata may be re-used in any medium without prior permission for not-for-profit purposes and re-sold commercially provided the OAI Identifier or a link to the original metadata record are given.*

Data policies generated through the tool tend to be more complex, but comprehensive. The Abertay Research Collection from the University of Abertay offers a very precise set of data access and reuse rules:[111]

> *2. Data reuse*
> *Policy for use of full-text and other full data items in the repository:*
> - *Anyone may access full items in all externally accesible Collections, apart individually embargoed items, free of charge.*
> - *Embargoed items are withheld from view due to legal requirements or to comply with publisher, funder or University policies.*
> - *Copies of open access full items generally can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided:*
>   - *the authors, title and full bibliographic details are given*
>   - *a hyperlink and/or URL are given for the original metadata page*
>   - *the original rights permission statement is given.*
> - *Full items must not be sold commercially in any format or medium without formal permission of the copyright holders.*
> - *Some full items are individually tagged with different rights permissions and conditions which must be adhered to.*

Interestingly, we were not able to find a single HEI repository using either the Open Data Commons licences, or the Open Government Licence. Lack of familiarity may be to blame, or perhaps those sites that have thought about intellectual property tend to use tools that are specifically designed for repositories. Whichever reason, there is a danger of the balkanization of UK data, with government, open data, and HEI repositories all using incompatible terms and conditions.

6.3 Contrasting HEI policies with other repositories

[109] An example of the excessive time and cost required to secure individual permission from each source in order to aggregate their content is detailed in Box 3 High Transaction Costs in McDonald D and Kelly U, *Intelligent Digital Options and The Value and Benefits of Text Mining*, JISC report (2012), http://bit.ly/TEpc9f.

[110] Policies can be found here: http://eprints.nottingham.ac.uk/policies.html.

[111] Terms of Use can be found here: http://is.gd/LiyBoX.

While it can be said that the policy landscape in HEIs seems to be continuously improving, it may be useful to contrast it with what is taking place with other types of repositories, as well as the practices regarding content mining in the proprietary scientific publication environment.

PubMed Central UK is typical of non-HEI and non-public sector repositories in the fact that it specifies that archived works may fall under full copyright protection, and therefore cannot be considered open access. In their copyright policy, they state:

> *Articles and other material in UKPMC usually include an explicit copyright statement. In the absence of a copyright statement, users should assume that standard copyright protection applies, unless the article contains an explicit statement to the contrary. In case of doubt, contact the journal publisher to verify the copyright status of an article.*

Similarly, PubMed Central UK has strong provisions against automated and systematic download of articles:

> *Crawlers and other automated processes may NOT be used to systematically retrieve batches of articles from the UKPMC web site. Bulk downloading of articles from the main UKPMC web site, in any way, is prohibited because of copyright restrictions.*

These restrictive practices seem to be the default outside of the open access publishing community. It is calculated that in the wider PubMed Central repository, 83% of all content is not licensed to allow content mining.[112] Similarly, high-profile academics and researchers have been publicly complaining about the difficulty of accessing published works for text mining purposes,[113] which has prompted the creation of the 3 principles of open content mining mentioned above.

In an interesting project, geneticists Max Haeussler and Casey Bergman started to document their attempts to obtain permission to text mine journal articles hosted by commercial scientific publishers and their repositories.[114] This negative response from Wolters-Kluwer is typical of the replies they are getting:

> *Any reproduction, distribution, performance, display, preparation of derivative works based upon, framing, capturing, harvesting, scraping, or collection of, or creating of hypertext or other links or connections to, any Site Materials or any other proprietary information of WKH, without WKH's advance written consent, is prohibited.*

The above seems to somewhat contradict research conducted by the Publishing Research Consortium (PRC), an industry association of academic publishers.[115] In the study the authors polled 190 journal publishers. Of these, 48% said that they had detected unauthorised crawling and downloads of their content, and 51% had received requests from individual research projects. 90% of those polled claimed that they had granted access for mining for research-focused mining requests, although 69% accepted that they dealt with requests on a case-by-case basis. This means that there is no wholesale, industry-wide approach to content mining, and proprietary "all rights reserved" copyright policies are the default position. There is clearly scope for improvement in this

---

[112] Nature "Editorial: Gold in the Text?" 483 *Nature* 124 (March 2012), http://bit.ly/Nx7c3M.

[113] Jha A, "Text mining: what do publishers have against this hi-tech research tool?" *The Guardian* (Wednesday 23 May 2012), http://bit.ly/Nx7GqD.

[114] Hosted at the UCSC Genome Bioinformatics Genocoding Project at http://text.soe.ucsc.edu/.

[115] Smit, E and Van Der Graaf M, "Journal Article Mining: The Scholarly Publishers' Perspective", 25:1 *Learned Publishing* 35 (2012).

area, and this could be the subject of future studies looking in more detail at a possible change in scientific academic publishing.

7. Conclusion

The objective of this paper has been to try to provide a clearer picture about the issues surrounding content mining in the UK's HEIs. Content mining offers great possibilities for academic research, so it is vital to provide investigators and management with enough information to help them make decisions about how to conduct and share their work. The present work has looked at content mining from two perspectives.

As users of content mining through research, there are still too many grey areas in the legal issues surrounding the subject. It seems safe to assume that at present the copying of text for analytical purposes may be considered infringing if one reads current case law in the most restrictive and conservative manner possible. Content mining does not fall easily into existing exceptions and limitations to copyright, and the scope of protection under the database right is also unclear. Even when done for research purposes, the scope of fair dealing for research and personal study is too narrow. If the law were changed to grant the exception for text mining in accordance to the recommendation contained in the Hargreaves Review of Intellectual Property, then the situation would become much clearer, which would help HEIs to conduct content mining in a safer legal environment and without the fear of infringing any existing rights.

Looking at HEIs as producers of minable research, the current licensing and reuse policies present in several institutional repositories are completely inadequate, and leave serious doubts with regards to reusability of data contained in the archives. Given the growing obligation for HEIs to make research publicly available, it would be expected that licensing and sharing policies would act accordingly and allow reuse for research purposes, but this is not the case at the moment. With regards to the licences used for open access and open data, there is a wealth of choice of open licences that may help to enable content mining. All of the three major suites cited in the paper can prove advantageous. However, too many choices may lead to licence incompatibility. The case study of the UK government's data hub offers a successful example where a top-down decision pertaining licensing choices resulted in high levels of adoption of one single licence. Whenever possible, standard licensing schemes should be encouraged to employ the top-down approach, which could result in licence harmonisation. This would produce in more compatibility between reuses, and a less fractured policy environment that would encourage content mining to a greater extent.

The reason why licensing choices and reuse policies are so important is that in an environment where more research is placed in the commons, this being a space where academic reuse is not only possible but encouraged, this would clearly obviate the first part of this study. The questions of whether content mining falls under existing fair dealing provisions, or whether there is permanent or temporary copying of data would be irrelevant if HEIs released most of their data using adequate reuse policies, or if they used an open access licence in their institutional archives. If a researcher mines data from a repository that allows reuse for research purposes, then there is no need to look at whether such mining infringes copyright, as the investigator has permission to run analytical tools on the content. Thankfully we seem to be headed towards such a future, but HEIs must do more with regards to their own policies.